PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 — A1

(11) International Publication Number: **WO 99/05614**

(43) International Publication Date: 4 February 1999 (04.02.99)

(21) International Application Number: PCT/IB98/01123

(22) International Filing Date: 23 July 1998 (23.07.98)

(30) Priority Data:
60/053,546 23 July 1997 (23.07.97) US

(71) Applicant *(for all designated States except US)*: DATOPS S.A. [FR/FR]; Maison des Professions Libérales, Allée N. Wiener, F–30000 Nîmes (FR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: GAY, Louis [FR/FR]; Datops, Maison des Professions Libérales, Allée N. Wiener, F–30000 Nîmes (FR). MASSIOT, Olivier [FR/FR]; Datops, Maison des Professions Libérales, Allée N. Wiener, F–30000 Nîmes (FR).

(74) Agents: MARTIN, Jean–Jacques et al.; Cabinet Regimbeau, 26, avenue Kléber, F–75116 Paris (FR).

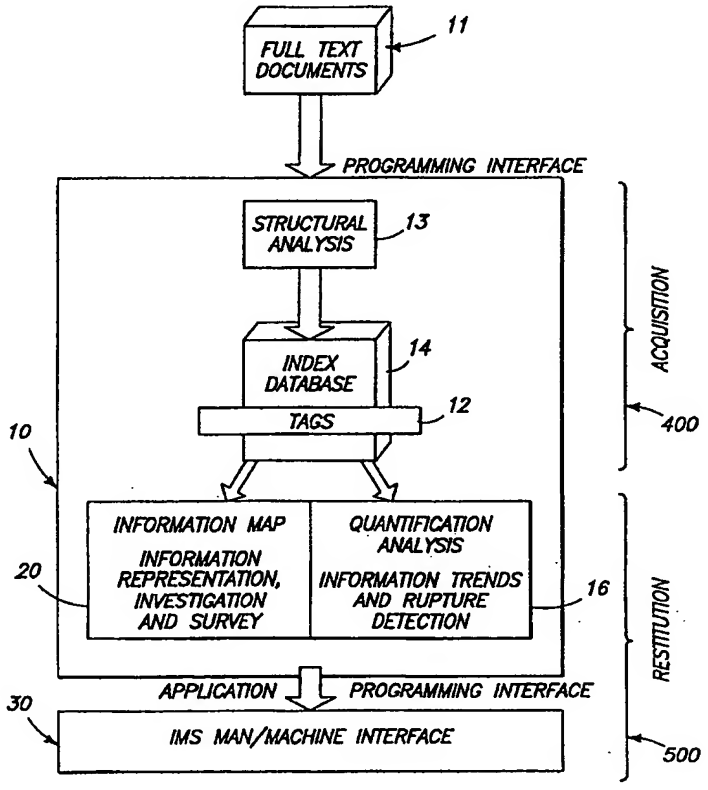
(81) Designated States: IL, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: INFORMATION MINING TOOL

(57) Abstract

An information mining tool comprising mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and means for determining parameters which relate to the evolution with time of said topics.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INFORMATION MINING TOOL

The present invention relates to an information mining technology which enhances the intelligence with which information can be analysed and in order to be better delivered to users.

TECHNICAL FIELD AND BACKGROUND OF THE INVENTION

Considering the huge number of information to which one can have access in particular through networks such as internet, there is a need for a processing tool permitting a fast evaluation of the information content of a huge number of collected text documents.

At the present time, the evaluation content of a set of collected documents is usually presented, in particular on the web, through the listing of the titles and possibly of summaries or of beginning sentences of these documents.

Such a listing does not permit to a user to clearly apprehend the informational content of the set of collected documents.

There is therefore a need for a data processing tool able to provide a synthetic presentation of the content of the collected documents, therefore offering fast reading capabilities.

Also, there is a need for an intelligent searching tool able to provide quantitative and qualitative analysis on a wide range of sources from structured to unstructured information.

In particular, there is a need for a tool permitting to follow the evolution with time of the informational content of a bank of collected documents and possibly to highlight for the user the modifications of the informational content of such a data tank which do not correspond to the normal evolution which could be expected.

Further, there is also a need for an intelligent searching tool adapting its processing of the information by revealing to the user topics which might be of concern for him, although not presented as so by the user in his queries.

SUMMARY OF THE INVENTION

The invention proposes a global system for:

.The Pull of Information : Using user programmable agents, which search through public and private information sources and retrieve relevant documents on a user-determined interval.

.The Mining of Information : Using complete Technologies for the Processing of Language as Text, as well as sophisticated signal and trend analysis, the invention analyses the retrieved documents, clusters them based on content, and matches them to each users unique information profiles. The information is prioritised for users based on relevancy of content, association with topics of interest, urgency and changeability.

.The Push of Information : Once information is analysed and processed to match unique user needs, it is delivered to the user in a variety of ways, including HTML page, mail pager and individual user reports.

This technology provides a number of capabilities, some of which appear to be differentiable today:

1. Ability to provide quantitative and qualitative analysis on a wide range of sources from structured to unstructured information.

In contrast to popular data mining approaches, the invention collects and analyses unstructured, qualitative information versus structured data. As a result, it can be used to analyse and profile information ranging from Web-based HTML pages, news delivery services to corporate text documents.

One of the key barriers to the growth and propagation of data mining tools is their requirement for a data warehouse, or structured information source. These warehouses are extremely expensive to create and maintain. The invention proposed can effectively "mine" information in its most natural form... a document. The analysis is not only qualitative, based on

processing of language, but also quantitative, particularly for determination of trends in information evolution.

2. Total information delivery solution.

The modularity of the system enables to use only one component of the technology : the Pull and Push portions of the product may be replaced with 3rd party products. However, the three components may be used to provide a total information delivery solution for specific application markets.

3. Profiling : User customisation and tuning.

Unlike popular information services, which provide pre-programmed topics of interest, the invention enables users to uniquely customise their information topics, based on their true area of interest. Additionally, the system actively monitors the users' work with the information, noting which information is "consumed" or not. Using this information, the system constantly tunes and updates the user's profile to provide more and more relevant information. As the users' profiles are permanently accurate, they can enter as parameters of each phase of information processing : to filter information (in the Pull), to process information and determine indicators according to users' needs (in Mining), to deliver information according to relevancy to user topics of interest (in the Push).

4. Fast Reading capabilities.

The analysis mechanisms used by the invention enable to provide information displays which graphically depict to a user the information's relevancy, proximity to other relevant topics, the intensity of the information, the trends surrounding the information and enable the user to dynamically change their views of the information. "Summaries" of the information content and direct access to original documents are available for each topic graphically displayed.

In other words, the mining tool proposed by the invention realises a Corporate Intelligent Channel.

To this end, the invention proposes an information mining tool comprising mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and means for determining parameters which relate to the evolution with time of said topics.

Such a tool provides quantitative as well as qualitative information.

1. Dynamic evolution

According to one important aspect of the invention, the mining tool comprise means to survey the time related evolution of a topic.

This permits to the mining tool to highlight the topics which become of importance for example in a given network, and in particular on the web or the news on internet.

In particular, this permits to detect discontinuities of evolution, even in case of topics corresponding to small signals.

2. Information Synthesis

The invention also proposes an information mining tool comprising mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and means to determine parameters characterising the relationship between topics, such as the average topological distance between the words corresponding to two topics or to the time related cross-similarity of two topics.

According to this important aspect of the invention, the mining comprise means to detect correlation between topics according to their time related evolution.

3. Cartographical display

In particular, the invention proposes an information mining tool which comprises push means which deliver to the user an information which relates to the topics, said push means comprising means to display on the screen of the user a map of the topics, said topics being presented in said map in form of nodes presented with links, the length of such a link between

two topics corresponding to the value of a parameter characterising the relationship between said topics.

4. Warnings detection

Advantageously, the push means comprise means to colour said nodes and links by using a colour code characterising the evolution with time of the topics and of their relationship parameters.

Such a presentation offers fast reading capabilities to the user.

5. Information Profiling

Further, the invention proposes an information mining tool comprising push means which deliver a set of topics and of corresponding documents in view of a particular query of the user and/or in view of a profiling file in which is stored a list of topics of interest for the user.

In particular, it comprises means for modifying the profiling file in view of the queries and/or selection of documents of the user.

Thus, the analysis of the documents base is recursive and takes into account the queries and fields of interest of the user, even though this evolution is not specifically formulated.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects and advantages of the invention will become more apparent and more readily appreciated from the following detailed description of the presently preferred exemplary embodiment of the invention taken in conjunction with the accompanying drawings, of which :

figure 1 and figure 2 are schematic drawings illustrating the architecture of the system ;

figure 3 illustrates an example of topics map to be displayed on the screen of the user.

DESCRIPTION OF THE INVENTION

As illustrated on figure 1, a system according to the invention is to process documents which might be collected from various sources.

These documents might be picked up in specific media such specific data bank servers, specific files or can be paper written documents electronically converted.

They can also be collected through the internet media by collecting mail messages or gathering documents from dedicated servers, or from the web or the news.

They can also be collected from intranet media by using information retrieval systems.

The documents are stored with corresponding metadata to constitute a text data base named «textual corpus», which can be processed by the information mining server which is illustrated on figure 2.

In the approach proposed, metadata is the key to scalability. Metadata is used to characterise the data for several purposes, including query processing, browsing and retrieval. Metadata may take different forms. It is required that metadata have the following properties : Effective (if the metadata says one information relates to other one, there is a great « probability » that is relevant), Concise (much smaller than the text it describes), Generated automatically (no human intervention required).

The information mining processing comprises two main tasks respectively hereinafter named acquisition and restitution.

In the acquisition processing, the textual corpus is processed to determine an index base which comprises a file of the topics representative of the informational content of the stored documents, as well as characteristics of these topics (hereinafter referred to as tags) and relationships which may exist between the topics. The index data base also comprises characteristics of documents (also called tags) and a file of indexation corresponding to a full text indexation of the documents.

The acquisition processing also uses a profiling base in which all the information relating to the profiles of the users are stored.

In the restitution step, the index base and the textual corpus are processed through an information retrieval processing and the informational

content of the documents can be displayed on the screen of the users in form of a schematic mapping of the topics.

ACQUISITION

The acquisition processing will be now described.

One aim of the acquisition processing is to add new tag values to documents, to extract related topics and relationships and also add tags to those topics and relationships.

This processing occurs on a working image in memory of the selected documents, these documents being kept stored in textual corpus in which they remain, without any change, during the whole processing.

1) In a first step, the text of each document is converted in series of a poor alphanumeric characters.

Such a conversion processing is for example of the type presented in :

- ABEL Y. "Indexation automatique de données textuelles", rapport de DEA "Contrôle des systèmes", Université de Technologie de Compiègne, Dépt Génie Informatique, Septembre 1993.

This conversion is made using an alphabet of 26 letters, 10 figures and 3 special characters (space, "." and "e").

The reading proceeding realises four processing :

- it transforms the alphabetical characters into capital letters with no accenting ;
- it keeps the numerical characters ;
- it replaces the other characters by e (thereby indicating that non interesting characters have been read) except for the characters space ("") and ".", these two characters being processed specifically. The space ("") indicates that two character strings are only separated by a gap and therefore are possibly part of a same sintagm. The "." permits to save the sets of initials, in which case it has not the same function as when it separates two sentences.

By the end of this first step, series of indexed alphanumeric characters are obtained, which are for example coded in UNICODE format.

2) In a second step, a lemma processing is performed. In particular, the texts can be processed to transform the verbs into their infinitive form, suppress detectable orthographic errors, detect the ambiguous words as well as the polysemic and homonymic and in such a case modify these words to suppress any ambiguity.

Far such a processing, one may refer to the following publications :

- PORTER, M.F. "An algorithm for suffix stripping", Program 14 (3), July 1980, pp. 130-137.

- CANDIDE N. "Acquisition automatique et polysémie en language naturel", Rapport de DEA, "Contrôle des systèmes", Université de Technologie de Compiègne, Dépt Génie Informatique, Septembre 1993, which is however specific to a text in French language.

3) After having been prepared in these first two steps, the textual documents are then structurally analysed.

3-1 In a first step of this structural analysis, a determination of the main language of each document is performed. To this end, a dictionary base is used which lists words which are the more representative of some languages. For example, a text incorporating a huge number of words such as «le», «la», «les» will be labelled as being mainly a French text.

The language which is determined is the language corresponding to the higher number of words of the dictionary base which appear in the processed text.

3-2 Then, in a second step, a parameter corresponding to an estimation of the structural complexity of the text is calculated. The parameter which is then provided permits to infer the kind of text to which the text processed belongs.

Such determination is for example based on neural network and uses a neural processing of a multilayer perceptron with a learning through a gradient retropagation algorithm. The topology of the network is of 41 input neurons, 10 output neurons and 15 to 20 neurons of an intermediate layer. The inputs of the input neurons are numerical characteristics which are calculated on the text through for example, for an HTML text its number of

images and their average length, the number and percentage of external links, the density of the text, etc..

The outputs neurons correspond to evaluations of the text complexity. The rate of success is around 95 % and varies with the nature of the corpus.

3-3 In a third step, a detection of the domain of the text is performed.

For example, it is detected whether the text analysed is a scientific text, a technical text or a business text.

The processing is performed by artificial neural networks, for example with a multilayer perceptron neural network using a retropagation learning algorithm and a topology with the same number of input neurons, output neurons or intermediate neurons as for the determination of the value of the structural complexity estimation parameter, the inputs of the input neurons being the same. The output neurons correspond to the various types of documents expected.

3-4 Then, in a fourth step, the text structures are detected. In this analysis, it is performed a structural syntaxic surface analysis which permits to detect in the text the series of alphanumeric characters which correspond to titles, sentences or paragraphs (patermatching processing).

3-5 In a fifth step, the texts are processed to perform a segmenting of their content into sentences, in the case where the punctuation is ambiguous. This segmenting is a non-trivial task, due to the ambiguity of many punctuation marks. The algorithm used is for example of the type described in :

PALMER D., «Tokenisation and Sentence Segmentation», The Natural Language Group, MITRE Corporation USA.

4) This structural analysis having been performed, the file obtained is then tokenised. The aim of this step is topic extraction.

This tokenising processing consists first in a statistical indexation of the words and in a particular in a calculation of the apparition frequency of words in the text.

The words are classified into hollow words, which are randomly distributed in the whole content of the file, (common language words with no

correlation with the topics of the texts) and sensible words which are not uniformly distributed in the file and mainly appear in some texts of the file.

This principle is enhanced by the exploitation a text preliminary classification by domains.

The method proposed rests on the idea that one can represent the importance of a term according to his number of occurrences and the number of domains where it appears.

Thus, the method uses the fact that a term "is diluted" on several domains or "is concentrated" on only one while bringing back the number of occurrences of a term to that of the domain where it appears more, when one wants to decrease the weight of the empty words, or with the sum of those where it appears less, when one wants to decrease the weight of the concepts.

More particularly, the hollow words are determined by selecting for each document the words which rate of occurrence is superior to a given threshold. A count corresponding to such a word is incremented each time this rate is superior to said threshold. The words selected as hollow words are those superior to a given selection threshold.

Having determined the hollow words, the file is processed to determine the topics.

This can be done through a lexicometry processing as proposed in :

ABEL Y. «Indexation automatique de données textuelles», rapport de DEA 'Contrôle des systèmes», Université de technologie de Compiègne, Dépt Génie Informatique, Septembre 1993.

In such a processing, the words following the hollow words are considered as potential sensible words. A count corresponding to said following words is incremented each time said word appears. The words selected as sensible words are those which occurrence rates in a text are superior to a given rate.

Once knowing the sensible words, one can determine the topics, by regrouping the sensible words which co-occur. To this end, it is examined for each word selected as a sensible word whether this word is followed by another sensible word separated from the first one by less than two hollow

words. If so, a count corresponding to the occurring of said couple of sensible words is incremented. And such a couple of sensible words is selected as corresponding to a single topic when the corresponding count is superior to a given threshold.

Afterwards, the complex words (attached words) are detected, as exposed in :

BOURIGAULT D. «Analyse syntaxique locale pour le repérage de termes complexes dans un texte», Xeme table ronde informatique & Egyptologie, Bordeaux 94,

as well as the private names, commercial names, or trademarks, which are to be treated independently.

The determined topics are stored in the index base (index data base 14 of Fig.2).

5) After having thus determined the topics, the process determines their related tags (12 of Fig. 2) in a fifth step.

5-1 For example, some of the stored tags 12 can correspond to all statistics information calculated in the previous steps of the processing, but now processed on each topic (all documents related to a topic).

5-2 Other tags can be classification tags describing the average type of activity (business, scientific, etc.) or the type of language (English or American expression, etc.) of a topic, these classification tags being determined through a neural network of the same type of the one used for the determination of the domain.

5-3 Other tags can also be trend parameters.

Trend parameters are processed by a «Dynamic Analysis » of information. The aims of this Dynamic Analysis are :

- the synthetic detection of the variation of information according to time,
- the measurement of the " informational risk " (or the risk by information) which can represent these tendencies.. The most obvious application of this Information Dynamic Analysis is to make it possible to warn a user of a dubious evolution of a subject, which enables him to be wary of the traditional models and to take safety measures.

To ensure determination of these warnings, it realise calculation of time series (hereinafter called « trends ») covering the impact on each theme of categories of influential events (political, industrial, financial...) »

Three types of trend parameters are for example preferred and advantageous.

A first trend parameter corresponds the number of documents in which the topic appears. It is hereinafter referred to as volumetric trend, which corresponds to the rough volume of published documents. It does not reflect really the intensity of the expressed opinion since it does not take account of the relevancy of documents toward the theme and of the number of sources or authors that expressed or retransmitted information. Volumetric trends can for example be compared for different periods of time in case of sets of documents collected from the same source with the same query agent.

A second trend parameter is the information intensity which corresponds to the ratio of another parameter called the global pertinency to the number of documents in the text.

The pertinency is a parameter determined for a given topic and a given text and corresponds to the number of apparition in said text of the words corresponding to said topic, with a ponderation attached to each of said words corresponding to the relevantness of said word relatively to said topic. The global pertinency of a topic corresponds to the sum of the pertinences for this topic of all the documents of the file.

The Pertinency can be :

- Calculated from the number of the occurring sought terms, as seen
- Or by the vector distance (in a space representation with N-dimensions) between the document and the theme or the class of concerned documents concerned. (It is here reminded that each document is usually represented in the documentary space by the vector corresponding to its lexical signature, the above-mentioned distance being the angle between two of said vectors).

The volume is balanced by the relevancy of the documents but can also balanced by the « Surface of publication » and by the size of the documents.

Surface of publication :

- in the case of the Newsgroup, information sources are subjected to the spams (misinformation operated by a same author by the massive diffusion of identical messages in one or more newsgroups).
- in the case of the press, it is necessary to be able to distinguish the integral recovery from a news which means propagation of information and the emission of new information or a new form of the same information.

The Surface of publication can then be defined as the number of authors divided by information volume.

The Information Intensity can be corrected by a multiplication with the following parameter :

Information Volume * (1 / log (Number of Volume))

A third trend parameter which is advantageously used is a signal value.

It can correspond to the difference between the value of the derivative with time of the volumic trend for a given query agent and the value of the derivative with time of the volumic trend for a larger reference query agent.

For example, a reference query can be «cows» whereas the specific query can be «mad cows».

As an alternative, when no reference query exists, the signal value may be determined as corresponding to the difference between the value of the derivative with time of the volumic trend for the given query and the average value in time of said derivative.

A fourth parameter which can then be used is the ratio between said signal and the reference or average volumetric derivative parameter.

By regularly calculating the value of this ratioparameter, one can detect the time at which the evolution of the information propagation breaks and therefore the time at which a topic becomes unexpectedly important.

5-4 Having determined these tags parameters, the texts are then processed to determine parameters characterizing the relationship between topics.

Relationships between topics are studied through the logical or semantic similarity between words detected from their context. The context of terms is represented as a set of attributes values illustrating :

. relative frequency of joint occurrence

. average topological distance

. time related cross-similarity

In essence, the model proposed assumes that the psychological similarity between two words is reflected in the way they co-occur in small subsamples of language and the way their evolutions through time are correlated. In other words the language produces words in a way that ensures an orderly stochastic mapping between semantic similarity and the calculated distance.

At first is determined a distance from co-occurring statistics : the square of the frequency of their joint occurrence in same documents related to the product of each occurrence.

One can also calculate the average topological distance between words corresponding to two topics. The topological distance is the likelihood of two words appearing in the same window of discourse – a phrase, a sentence, a paragraph... This distance is inversely related to their semantic distance, that is directly related to their semantic similarity. Observing the relative frequency of their joint occurrence in such windows is a part of the estimation of the relative similarity of any pair of words.

One important feature compared to the standard « similarity » or « distance » measures proposed in the information retrieval literature, is the determination of the cross correlation parameter which is a cross correlating calculation with time of the intensity or volumetric trend parameters of two topics.

There is two ways of process cross-similarity or cross-correlation.

1/ First, the standard Autocorrelation function is the Fourier transform of power spectrum :

$g(\tau) = \langle x(t)\ y(t+\tau) \rangle$ (< ... > denotes time average) . x and y being information intensity or volumetric parameter of the two topics.

A highly positive auto-correlation value means that correlation is established between the elements of X and those of Y temporally shifted. A minus coefficient implies an anti-correlation. It is relatively difficult to interpret such a test. A signal does not consist of only one and single periodicity and in general this periodicity is not even constant in amplitude and/or time.

Correlation as a measure is dependant on a Gaussian distribution, which we know not to hold for the data that we want to analyse.

2/ A better indicator can be found in Mutual Information.

Mutual information makes no assumption about the distribution of the measured series, and is therefore the most attractive measure to hand.

Mutual Information is a concept conceived by Claude Shannon (1949). Mutual Information attempts to measure in bits the amount of information that can be inferred about one series of symbols by another. A derivation of this concept is used. In general given two series x and y with indexes i and j respectively, the average mutual information I(x,y) can be calculated as:

$$I(x,y) = \sum_{i,j} P(x_i, y_j) \log\left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right)$$

Mutual information is positive and symmetrical ($I(x,y) > 0$ and $I(x,y) = I(y,x)$). The maximum and the first minimum in the resulting curve of MI against n are significant values.

The mutual information for the given series at increasing offsets is calculated. At each stage the minimum and maximum values so far are calculated. If the current offset is the new low the maximum value is reset to that of the new low. When the processing is finished, that is that all offsets up to 17 have been tried, the point at which the low occurred is examined. If this is the last offset then the MI graph was continually decreasing, and a minimum can not be found. If not, the subsequent maximum is inspected. If

this exceeds 1.1 times the minimum, then a winner is declared, and the offset at which that minimum occurred is selected as the separation distance, otherwise it is assumed that a minimum can not be found.

One can also calculate an « informational distance » mixing time related correlation, co-occurring statistics and topological distances.
The determination of the clusters – sets of topics which share neighbour relations – can be based on the informational distance calculated like above. One can fits all the pairwise similarities into a common space of high dimensionality and apply on it clustering and classification.
The construction of the map (described below) exploits the informational distance to represent the neighbour nodes related to the central node of a map.

RESTITUTION

Following the acquisition step 400, information are then restituted in restitution step 500 (Fig. 2).

An information retrieval processing is advantageously performed, as further described below.

The queries can be factual or/and boolean, the information retrieval being then performed on the topics file and on the indexation file of the index base.

One can also use queries of natural language, in which case the queries formulated are transformed into boolean queries by a program operating on the index base files.

The selected documents can be classified by order of pertinence regarding the formulated query. They also can be highlighted by the values of trend parameters which denote an abnormal evolution of the topic.

The user having chosen a pertinency level, the system displays on the screen of the user the topics which appear in the selected documents having pertinence superior to said pertinence level.

The pertinence level of a topic is determined by using tags parameters.

For example, the system can display a map in which the topics appear in form of nodes distributed on the screen (see figure 3).

These nodes are represented with links in the case of topics having a high cross-correlating parameter or a topological or informational distance superior to a given threshold.

In such a case, the lengths of the links correspond or tend to correspond to the topological or informational distances between said topics. In case too much nodes are linked to each other to have an exact representation of the distance between one another, the system will optimise the repartition of the nodes in order to minimise the difference between the distance from a node to another.

In case two topics, having a short informational distance or more specifically a high time related cross-similarity, the nodes of these two topics can be merged in a single node.

Possibly, these links and nodes are colorated for taking into account their evolution in the last period of time.

For example, a node will be colorated in red in case one of its trend parameter is highly increasing (for example if a break of propagation is detected).

It is colourated in blue in case its trend parameter is decreasing in the last period.

Also, the links will be colourated to take into account the evolution of the informational distance or more specifically the cross-similarity parameter to which they correspond.

Further, when the user clicks on the node of a topic to select it, the process gives him a list of the documents concerned by third topic with a pertinency hierarchy.

And for each document, the user is provided with a summary which corresponds to sentences where words corresponding to the topics are found.

Also, for a given query of the user, the system can display tags of the topics corresponding to this query and can also determine new tags which are specific to the query.

For example, the determined trends can be displayed to the user in form of graphs giving their value with time.

RECURSIVITY

The system presents a processing by which it takes into account the behaviour of the user.

For example, the system memorises a profile of the user in which are stored topics of interest for him.

Such a profiling can comprise a structural part and a personal part, as well as an implicit part.

The structural part comprises topics which are of interest for the environment of the user (for example, topics concerning his company). It is divided into inalienable topics - which in any case appear in the mapping display of the user - and dynamic topics - which are selected by the user himself.

The personal part comprises topics which relate to the user and not to his environment (for example his own fields of interest in the company). It is also divided into inalienable topics and dynamic topics.

The implicit part of the profiling comprises topics which in time appear to be cross-correlated with topics of the structural or personal part of the profiling.

New topics created by the user himself are defined through the formulation of a new query. This new topic will be characterised in the profiling file by the words and expressions corresponding to the query.

These new topics can also be selected by the user in the file of topics of the index base.

Each time a particular document is read by the user during a time superior to a given threshold, the document is attached to the user profile. These documents can be processed as same as the processed of collected documents has been defined. Topics can be extracted so that the pertinence of the corresponding topics is also increased.

It will be understood that the processing here above described may be reduced to tangible apparatus components primarily by programming the method steps into computer programs, and installing the computer programs on computer or machine readable elements such as ROM or RAM, hard drive, compact disk, tape, diskette, cassette, or other tangible program receiving media. The programs which are stored on tangible media may then be assembled into a complete apparatus by installing the programs onto or into one or more computers which may be linked together by known methods for linking computers, such as networking or other cabling means, including by telecommunications systems.

What we claim is

1. An information mining tool comprising mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and means for determining parameters which relate to the evolution with time of said topics.

2. An information mining tool as in claim 1, wherein the mining means comprise means to determine a volumic topic trend parameter of a topic, said trend parameter corresponding to the number of documents in which said topic appears.

3. An information mining tool as in claim 1, wherein the mining means comprise means to determine for a given topic and a given text, the number of apparition in said text of the words corresponding to said topic, with a ponderation attached to each of said words corresponding to the relevantness of said word relatively to said topic.

4. An information mining tool as in claim 3, wherein the mining means comprise means to determine the global pertinency of a topic, said global pertinency corresponding to the sum of the pertinencies for this topic of all the documents of the documents base.

5. An information mining tool as in claim 4, wherein the mining means comprise means to determine an informational intensity topic trend parameter, said parameter corresponding to the ratio of the global pertinency of the topic to the number of documents in the documents base.

6. An information mining tool as in claim 5, wherein the mining means comprise means to determine a signal value corresponding to the difference between the value of the derivative with time of the volumic trend for a given query agent and either the average value in time of said derivative or the

value of the derivative with time of the volumic trend for a larger reference query agent.

7. An information mining tool as in claim 6, wherein the mining means comprise means to determine the ratio between said signal and either the average value in time of said derivative or the value of the derivative with time of the volumic trend for a larger reference query agent.

8. An information mining tool according to the preceding claims, characterising in that it comprises mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and means to determine parameters characterising the relationship between topics.

9. An information mining tool as in claim 8, wherein the mining means comprise means to determine the average topological distance between the words corresponding to two topics.

10. An information mining tool as in claim 8, wherein the mining means comprise means to determine a volumic topic trend parameter of a topic, said trend parameter corresponding to the number of documents in which said topic appears.

11. An information mining tool as in claim 8, wherein the mining means comprise means to determine for a given topic and a given text, the number of apparition in said text of the words corresponding to said topic, with a ponderation attached to each of said words corresponding to the relevantness of said word relatively to said topic.

12. An information mining tool as in claim 11, wherein the mining means comprise means to determine the global pertinency of a topic, said global pertinency corresponding to the sum of the pertinencies for this topic of all the documents of the documents base.

13. An information mining tool as in claim 12, wherein the mining means comprise means to determine an informational intensity topic trend parameter, said parameter corresponding to the ratio of the global pertinency of the topic to the number of documents in the documents base.

14. An information mining tool as in claim 10, wherein the mining means comprise means to determine at least a parameter corresponding to the cross-correlation of the values in time of the volumic topic trend parameter of two topics.

15. An information mining tool as in claim 13, wherein the mining means comprise means to determine at least a parameter corresponding to the cross-correlation of the values in time of the informational intensity topic trend parameter of two topics.

16. An information mining tool as in claim 14 or 15, wherein said parameter is determined by an auto-correlation calculator.

17. An information mining tool as in claim 14 or 15, wherein said parameter is determined by an average mutual information calculation.

18. An information mining tool as in claim 8, wherein it comprises push means which deliver to the user an information which relates to the topics, said push means comprising means to display on the screen of the user a map of the topics, said topics being presented in said map in form of nodes presented with links, the length of such a link between two topics corresponding to the value of a parameter characterising the relationship between said topics.

19. An information mining tool as in claim 18, wherein the push means comprise means to colour said nodes and links by using a colour

code characterising the evolution with time of the topics and of their relationship parameters.

20. An information mining tool according to any of the preceding claims, comprising mining means for processing documents stored in a data base in order to extract the topics to which these documents relate and push means which process a file in which said topics are stored to retrieve a set of topics and of corresponding documents in view of a particular query of the user and/or in view of a profiling file in which is stored a list of topics of interest for the user.

21. An information mining tool as in claim 20, wherein it comprises means for modifying the profiling file in view of the queries and/or selection of documents of the user.

22. An information mining tool as in claim 21, wherein the profiling file comprises a structural part comprises which comprises topics which are of interest for the environment of the user, a personal part which comprises topics which specifically relate to the user and an implicit part which comprises topics which in time appear to be correlated with topics of the structural or personal part.

23. An information mining tool as in claim 22, wherein the structural part or personal part of the profiling file are divided into inalienable topics, which are imposed to the user and dynamical topics, which are chosen by the user.

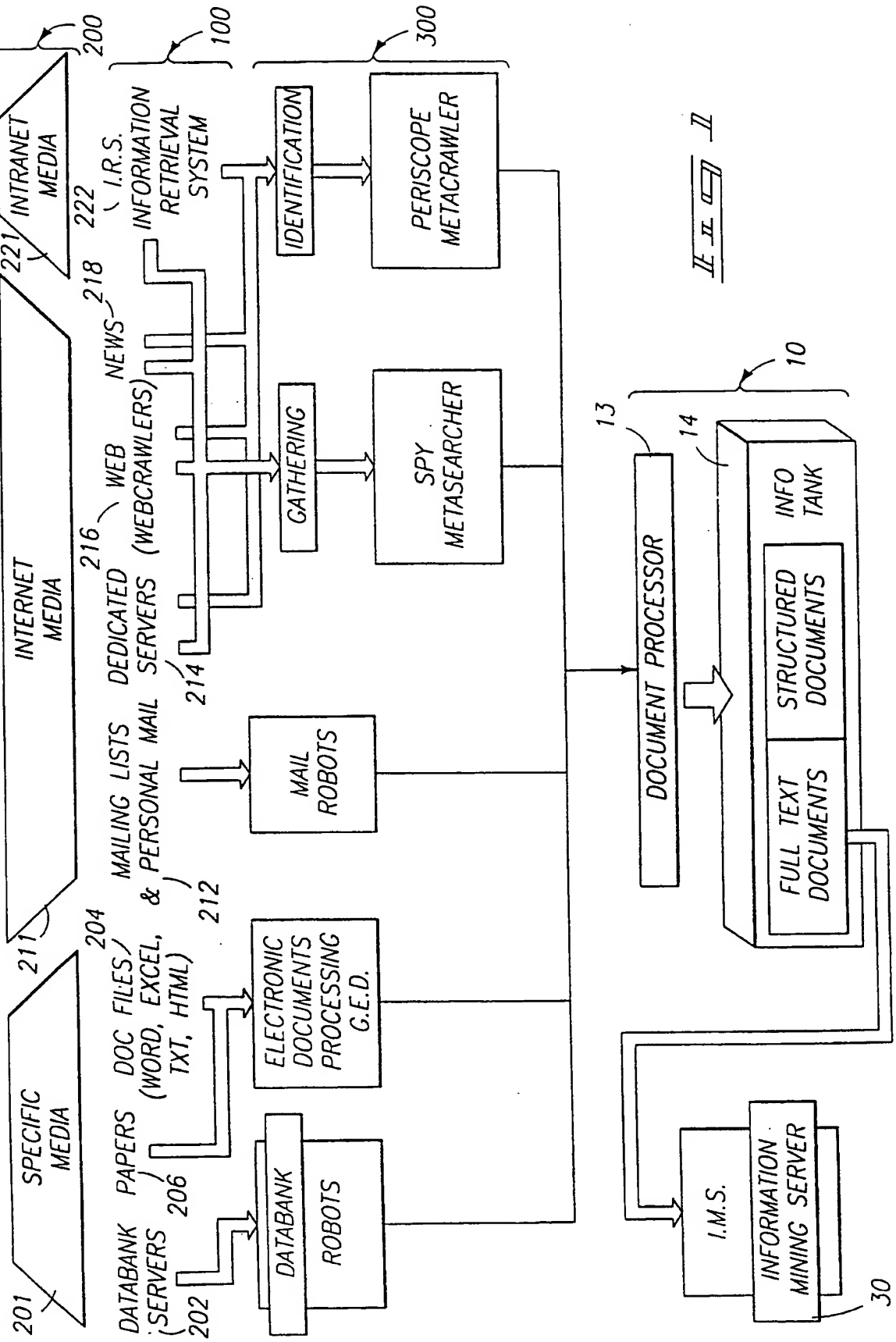

200
100
300
INTRANET MEDIA
I.R.S. INFORMATION RETRIEVAL SYSTEM
IDENTIFICATION
PERISCOPE METACRAWLER
221
222
218
NEWS
INTERNET MEDIA
216
WEB (WEBCRAWLERS)
GATHERING
SPY METASEARCHER
DEDICATED SERVERS
214
MAILING LISTS & PERSONAL MAIL
212
MAIL ROBOTS
13
DOCUMENT PROCESSOR
14
10
INFO TANK
STRUCTURED DOCUMENTS
FULL TEXT DOCUMENTS
211
204
SPECIFIC MEDIA
DOC FILES (WORD, EXCEL, TXT, HTML)
ELECTRONIC DOCUMENTS PROCESSING G.E.D.
PAPERS
206
DATABANK (SERVERS)
202
DATABANK
ROBOTS
201
I.M.S.
INFORMATION MINING SERVER
30

FIG II

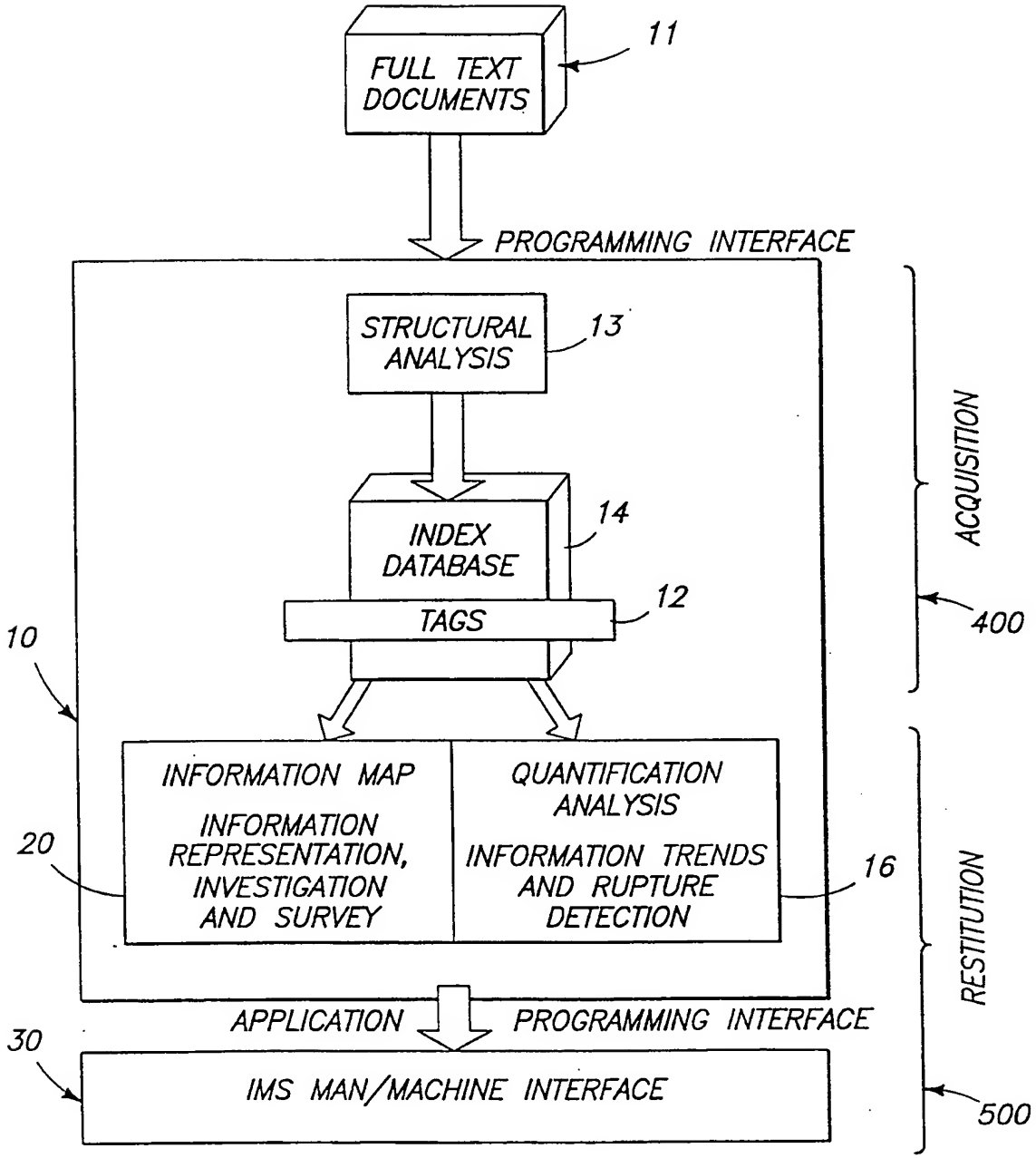

FULL TEXT DOCUMENTS
11
PROGRAMMING INTERFACE
STRUCTURAL ANALYSIS
13
INDEX DATABASE
14
TAGS
12
ACQUISITION
400
10
INFORMATION MAP
INFORMATION REPRESENTATION, INVESTIGATION AND SURVEY
20
QUANTIFICATION ANALYSIS
INFORMATION TRENDS AND RUPTURE DETECTION
16
RESTITUTION
APPLICATION PROGRAMMING INTERFACE
30
IMS MAN/MACHINE INTERFACE
500

Fig 2

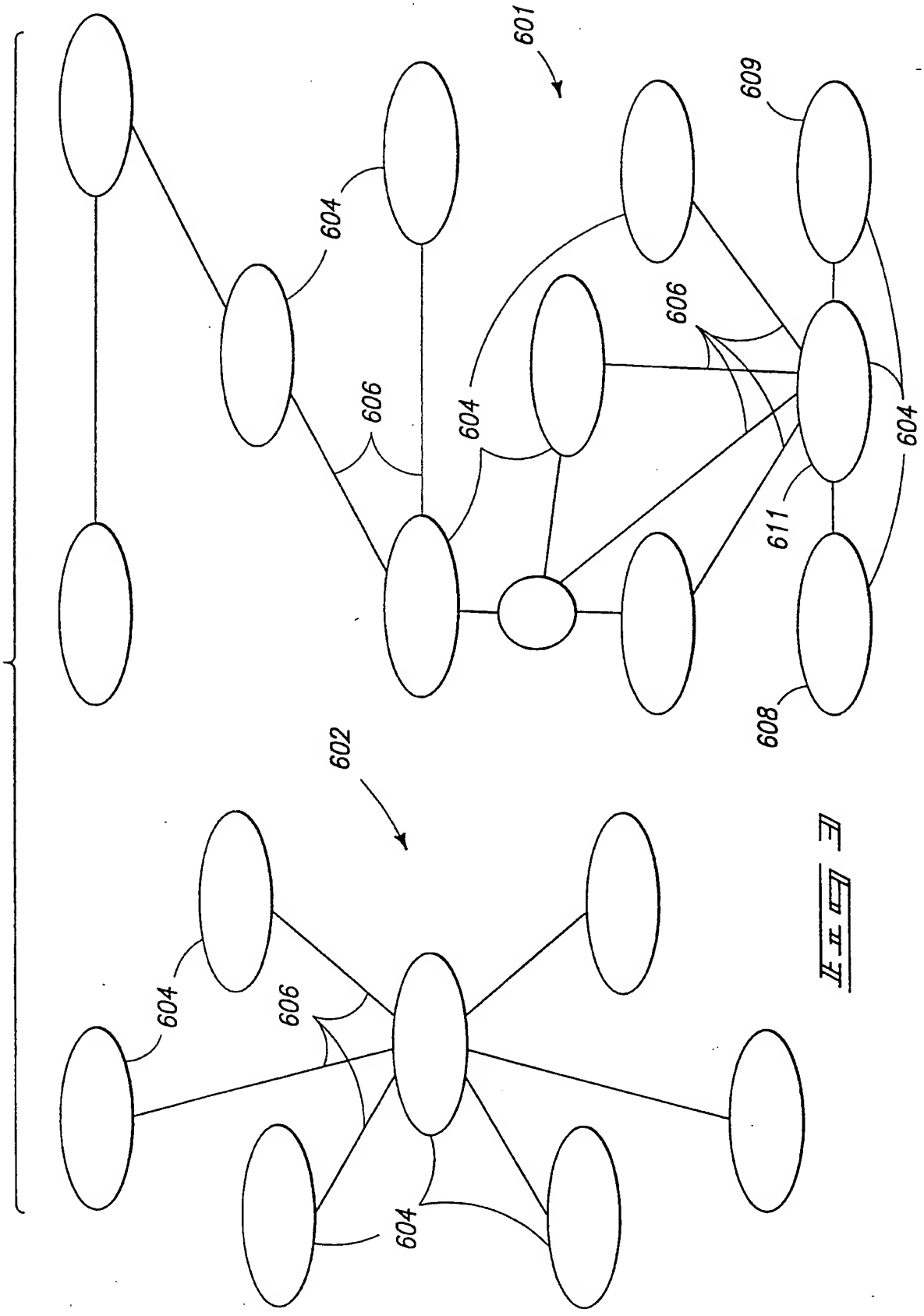


Fig 3

INTERNATIONAL SEARCH REPORT

In tional Application No
PCT/IB 98/01123

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LE MONDE INFORMATIQUE, no. 705, 17 January 1997, pages 1-11, XP002082836 http://www.lmi.fr/705/705p22.html see column 3, line 1 - line 7 see column 5, line 1 - column 11, line 52 ---	1
A	CHEN H ET AL: "INTERNET CATEGORIZATION AND SEARCH: A SELF-ORGANIZING APPROACH" JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION, vol. 7, no. 1, March 1996, pages 88-102, XP000619822 see page 92, right-hand column, paragraph 4 - page 93, right-hand column, paragraph 4.2; figures 1-3 see page 95, left-hand column, paragraph 4.2.2 - paragraph 5 ---	1-3, 8-11,18, 20

-/--

[X] Further documents are listed in the continuation of box C. [ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 October 1998	12/11/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Fournier, C

2

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

In tional Application No
PCT/IB 98/01123

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	GINSBERG A: "A UNIFIED APPROACH TO AUTOMATIC INDEXING AND INFORMATION RETRIEVAL" IEEE EXPERT, vol. 8, no. 5, 1 October 1993, pages 46-56, XP000413472 see the whole document ---	1-23
A	WONG J W T ET AL: "ACTION: AUTOMATIC CLASSIFICATION FOR FULL-TEXT DOCUMENTS" SIGIR FORUM, vol. 30, no. 1, 21 March 1996, pages 26-41, XP000699962 see abstract -----	1,3,8,11

2